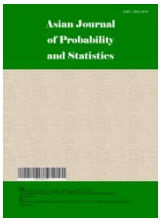

Asian Journal of Probability and Statistics

12(2): 41-51, 2021; Article no.AJPAS.67176
ISSN: 2582-0230

A New Calibration Estimator of Population Mean for Small Area with Nonresponse

J. Iseh Matthew[1*] and J. Bassey Kufre[2]

[1]Department of Statistics, Akwa Ibom State University, Mkpat Enin, Nigeria.
[2]Statistics Department, Central Bank of Nigeria, Abuja, Nigeria.

Authors' contributions

This work was conceptualized and carried out by author JIM. in collaboration with author JBK. The first draft of the manuscript was written by author JIM, including the protocol. Author JBK reviewed the manuscript and managed the structure of the manuscript. Both authors evaluated the revised manuscript, discussed the results, and finalized the paper.

Article Information

DOI: 10.9734/AJPAS/2021/v12i230286
Editor(s):
(1) Dr. Dariusz Jacek Jakóbczak, Koszalin University of Technology, Poland.
Reviewers:
(1) Corina Flores Hernandez, University of Guanajuato, Mexico.
(2) Traore Diakalya, Université Joseph Ki-Zerbo, Burkina Faso.
Complete Peer review History: http://www.sdiarticle4.com/review-history/67176

Original Research Article

Received 06 February 2021
Accepted 12 April 2021
Published 20 April 2021

Abstract

This paper considered the challenges of population mean estimation in small area that is characterized by small or no sample size in the presence of unit nonresponse and presents a calibration estimator that produces reliable estimates under stratified random sampling from a class of synthetic estimators using calibration approach with alternative distance measure. Examining the proposed estimator relatively with existing ones under three distributional assumptions: normal, gamma, and exponential distributions with percent average absolute relative bias, percent average coefficient of variation, and average mean squared error as evaluation criteria using simulation analysis technique, the new estimator exhibited a more reliable estimate of the mean with less bias and greater gain in efficiency. Further evaluation using coefficient of variation under varying nonresponse rates to validate the results of variations suggests that the estimator is a suitable alternative for small area estimation. This finding has therefore contributed to the development of an ultimate estimator for small area estimation in the presence of unit nonresponse.

[1]Department of Statistics, Akwa Ibom State University, Mkpat Enin, Nigeria: Email- eeseaglechild@gmail.com
[2]Statistics Department, Central Bank of Nigeria, Abuja: Email- kjbassey@gmail.com;

**Corresponding author: Email: eeseaglechild@gmail.com;*

Keywords: Calibration; distance measure; nonresponse; small area estimation; synthetic estimators.

JEL CLASSIFICATION: 62D05.

1 Introduction

The use of synthetic estimators in small area estimation (SAE) has become one popular technique in small area estimation. This is so because it could produce reliable estimates when there are small or no sample observation in areas of interest. This was first examined by [1] in the public health service of the United States of America. The essential property of the synthetic estimators made it so attractive in SAE, unlike the direct estimators that are based on the sample information obtained from the area of interest, which is not reliable due to lack of effective sample size in areas of interest. This indirect method of estimation was also applied in the estimation of a mean income of a family, the average production of crops in blocks, and the number of unemployed persons in councils, among others.

Though synthetic estimation technique has been adopted by different authors to compensate for the challenges of small sample sizes in SAE [2-5], with other contributors using calibration weights as a means of improving the precision, ([6-19]), small or no sample size problem in the presence of unit nonresponse remains a gap in the literature.

Earlier introduction of a calibration estimation approach by [20] included a distance function to account for auxiliary information in the estimation, a method often refers to as *"creating estimators by benchmarking the auxiliary information to external controls".* Thereafter, [7] and [8] posited that the proposed distance measure was not effective in addressing the dual problem of small sample size and nonresponse in a domain of interest. A call for an estimator that could address this dual problem by [21] led to a study by [22] that showed that calibration estimators performed poorly when sample sizes become very small but more efficient as sample size increased, whereas, synthetic estimator becomes more effective at domains with small sample sizes. Having in mind the report from this study and in consideration of the proposition by [23] on the choice of alternative distance measure for calibration weight to bridge the gap between the original design weight and the calibration weight under nonresponse. The objective of this paper, therefore, is to formulate an alternative calibration estimator for small area in the presence of unit nonresponse.

Thus, the paper proposes an alternative synthetic estimator that addresses the limitations in the previous studies using the calibration approach by adopting the procedures of [7,8,23].

1.1 Theoretical underpinning

Consider a finite population consisting of N units which are divided into D non-overlapping domains $U_d,\ d = 1,2,\dots,D$ with N_d units such that $\sum_d^D N_d = N$. Let the population be further partitioned into G non-overlapping groups (considered to be strata) which are considered to be larger than the domains $U_{.g},\ g = 1,2,\dots,G$ with $N_{.g}$ (the population size of the group), such that $\sum_g^G N_{.g} = N$, so that the G groups cuts across the D domains to form a grid of DG cells denoted by U_{dg} with N_{dg} units, then $U = \bigcup_{d=1}^D U_d = \bigcup_{g=1}^G U_{.g} = \bigcup_{d=1}^D \bigcup_{g=1}^G U_{dg}$ and $N = \sum_d^D N_d = \sum_g^G N_{.g} = \sum_d^D \sum_g^G N_{dg}$. The sample s is analogously partitioned into domain subsamples s_d, group subsamples $s_{.g}$ and cells subsamples s_{dg} with corresponding sample sizes $n,\ n_d,\ n_{.g}$ and n_{dg} as $s = \bigcup_{d=1}^D s_d = \bigcup_{g=1}^G s_{.g} = \bigcup_{d=1}^D \bigcup_{g=1}^G s_{dg}$ and $n = \sum_d^D n_d = \sum_g^G n_{.g} = \sum_d^D \sum_g^G n_{dg}$. The cells subsamples n_{dg} are assumed to be random. Ordinarily, n_d and $n_{.g}$ are also random but $n_{.g}$would be fixed if the g^{th} group is a stratum from which a fixed number of elements is drawn. Let Y be the study variable which values y_{dgk} are known for just the element of a sample s, where $k = 1,2,\dots,N_{dg}$ (the number of population units in the $(dg)^{th}$ cell) and X be the auxiliary variable which values x_{dgk} may or may not be known *apriori* for all units in U.

For different reasons, there are missing units in the sample, s and is considered as the only source of data contamination in this work. If we further denote the response set by r, and instead of the original sample size, n, we receive an incomplete response to n_r, then, the response probability $P(k \in r | k \in s)$, where $r \subset s$, is a responded sample, $r_d = r \cap s_d,\ r_{.g} = r \cap s_{.g}$ and $r_{dg} = r \cap s_{dg}$ with their respective sizes n_{r_d}, $n_{r_{.g}}$, and $n_{r_{dg}}$. Now, let us consider the following estimators for domain estimation in the presence of unit nonresponse:

a. domain estimator of a population mean in the presence of unit nonresponse (direct estimation):

An estimator under nonresponse in estimating population total was suggested by [4]. In a follow-up, an estimator for estimating domain population mean $\bar{Y}_d = \frac{1}{N_d}\sum_{g=1}^{G}\sum_{k\in U_{dg}}^{N_{dg}} Y_{dgk}$, $d = 1, \dots, D$ is obtained as:

$$\hat{\bar{y}}_{dr} = \frac{1}{N_d}\sum_{g\in r_{.g}}^{G}\sum_{k\in r_{dg}}^{N_{dg}} \frac{y_{dgk}}{\pi_k \hat{\theta}_k} \tag{1}$$

where $\hat{\theta}_k$ is the estimate of the unknown response probability θ_k. Equation (1) is an extension of the basic Horvitz-Thompson estimator to a selection in two phases.

b. calibration estimator of a population mean in the presence of unit nonresponse (calibration approach):

Lundstrom and Sarndal [7,8] proposed a single step weighting scheme through calibration approach as an improvement to Equation (1) for estimating the domain population mean in the presence of nonresponse as:

$$\hat{\bar{y}}_d = \sum_{r.g}^{G} P_{dg}\, \bar{y}_{dg} \tag{2}$$

where $\bar{y}_{dg} = \frac{1}{n_{r_{dg}}}\sum_{k\in r_{dg}}^{N_{dg}} y_{dgk}$ is the sample mean for the response set in the $(dg)^{th}$ cell and P_{dg} is the calibration weights formed to be as 'close as possible' to the basic stratum weights W_{dg} in stratified random sampling at the two levels of information which satisfy the calibration equations.

When information is available at the population level of the auxiliary variable, the calibration weight becomes $P_{dg} = W_{dg} v_{kU}$ and the calibration estimator in equation (2) becomes

$$\hat{\bar{y}}_{dU} = \sum_{r.g}^{G} W_{dg} v_{kU}\, \bar{y}_{dg} \tag{3}$$

where $v_{kU} = 1 + q_{dg}\left(\sum_{U.g}^{G}\bar{x}_{dg} - \sum_{r.g}^{G} W_{dg}\bar{x}_{dg}\right)\left(\sum_{r.g}^{G} W_{dg} q_{dg}\bar{x}_{dg}^2\right)^{-1}\bar{x}_{dg}$ and $\bar{x}_{dg}$ is the auxiliary variable analogously defined as $\bar{y}_{dg}$. Equation (1) was compared to Equation (3) by assuming that $v_{kU} = \frac{1}{\hat{\theta}_k}$, where $\hat{\theta}_k$ is the estimate of the unknown θ_k. It should be noted that the idea of calibration weights as applied by [7,8] in Eq. 3, has lessened the burden of finding the unknown response probabilities θ_k in the two-phase estimator of Eq. 1, by [4].

Again, when information on the population of the auxiliary variable is unknown, calibration is done on the sample estimates, and the weight becomes $P_{dg} = W_{dg} v_{ks}$ and the calibration estimator in Eq. 2 is obtained as:

$$\hat{\bar{y}}_{ds} = \sum_{r.g} W_{dg} v_{ks}\, \bar{y}_{dg} \tag{4}$$

Where $v_{ks} = 1 + q_{dg}\left(\sum_{s.g}^{G} W_{dg}\bar{x}_{dg} - \sum_{r.g}^{G} W_{dg}\bar{x}_{dg}\right)\left(\sum_{r.g}^{G} W_{dg} q_{dg}\bar{x}_{dg}^2\right)^{-1}\bar{x}_{dg}$ and q_{dg} is known as the tuning parameter and it is assumed to be '1' in this work.
The estimators in equations (3) and (4) are very unstable with small sample size. More so when the domain of interest has no sample unit it becomes difficult (if not impossible) to be computed given that they are modified direct estimators.

2 Methodology

This study is designed to proffer solution to inherent challenges in small area estimation, which involved a combination of small/no sample size, and nonresponse adjustment. As suggested by [21], one should not be tackled in isolation of the other.

In what followed, bivariate observations, (x_{ij}, y_{ij}), were first generated, which comprised a finite population of size 4950 units. The population U considered was created by generating data for three separate subsets of the

populations termed *groups* (strata) with different intercepts and slopes. Each group was split into ten domains that are mutually exclusive and exhaustive as follows: *Group 1;* $U_{11}, U_{21}, \ldots, U_{101}$, *Group 2;* $U_{12}, U_{22}, \ldots, U_{102}$, and *Group 3;* $U_{13}, U_{23} \ldots, U_{103}$. The number of units in each cell N_{dg} were sequentially allocated in a monotonic manner: cell U_{11} with 20 units; cell U_{21} with 30 units; and cell U_{103} with 310 units. The values of x in each group were generated from three different distributions, *Gamma* $(\alpha = 10, \beta = 1)$, *Norm* (5,1), and *Exp* (0.5) distributions. The simulation for the variable of interest y was obtained using the model:
$y_{dk} = \beta_{og} + \beta_{1g} x_{dk} + v_d + e_{dk}$, where $d = 1,2,\ldots,30;\ k = 1,2,\ldots,N$ and $g = 1,2,3$; $e_{dk} \sim N(0, C_{dk}^2 \sigma_e^2)$, $v_d \sim N(0, \sigma_v^2)$. It is assumed that $\sigma_e^2 = \sigma_v^2 = 20^2 = 400$ for the gamma distribution, $\sigma_e^2 = \sigma_v^2 = 1^2 = 1$ for normal and exponential distributions. $c_{dk} = x_{dk}$ is set to reflect the heterogeneity of the model errors for the synthetic and calibration estimators.

Secondly, the sample size settings generated from the above model were pegged at 5%, 10%, 20%, and 25% to reflect the case of small area and then, the use of exponential response probability; $\theta_k = 1 - e^{-CX_k}$, $k \in U$, where c is chosen according to the desired average response probability. This study adopts values varying between 0.60, 0.70, and 0.86 (the latter value being the chosen response probability in [9])

2.1 Summary Statistics

Summary statistics of the simulated data can be obtained using Average Percent Absolute Relative Bias, Average Mean Square Error and Average Percent Coefficient of Variation $\%\overline{ARB}$, $\overline{MSE}$ and $\%\overline{CV}$ respectively and are obtained as:

$$\%\overline{ARB}(\hat{\bar{y}}_{dP}) = \left[\frac{1}{D}\sum_{d=1}^{D} ARB(\hat{\bar{y}}_{dP})\right] \times 100, \text{ where } ARB(\hat{\bar{y}}_{dP}) = \left|\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\bar{y}}_{dP}^{(r)}}{\bar{Y}_d} - 1\right)\right|$$

$$\overline{MSE}(\hat{\bar{y}}_{dP}) = \frac{1}{D}\sum_{d=1}^{D} MSE(\hat{\bar{y}}_{dP}) \text{ and } MSE(\hat{\bar{y}}_{dP}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{y}}_{dp}^{(r)} - \bar{Y}_d\right)^2, and$$

$$\%\overline{CV}(\hat{\bar{y}}_{dP}) = \left[\frac{1}{D}\sum_{d=1}^{D} CV(\hat{\bar{y}}_{dP})\right] \times 100, \text{ where } CV(\hat{\bar{y}}_{dP}) = \frac{\sqrt{MSE(\hat{\bar{y}}_{dP})}}{\bar{Y}_d} \quad ,$$

where $\hat{\bar{y}}_{dP}^{(r)}$ and $\hat{\bar{y}}_{dE}^{(r)}$ denote say, the proposed and existing estimators, respectively, produced for the r^{th} sample, $r = 1,2,\ldots,R$, and for each small area $d = 1,2,\ldots,D$. For each selected sample in each simulation run, $r = 1,2,\ldots,R$ $(R = 100{,}000)$, we shall compute estimates of $\overline{Y}_d$ for the estimators.

Remark: In small area estimation, Molina and Rao (2010) suggested a benchmark value for $\%\overline{CV}(\hat{\bar{y}}_{dP})$ at 20-25% as being reliable. As a result, high value of $\%\overline{CV}(\hat{\bar{y}}_{dP})$ above 25% is considered as unreliable estimates while estimators with values of $\%\overline{CV}(\hat{\bar{y}}_{dP})$ below 25% is considered reliable and suitable for SAE.

2.2 Calibration estimator for small area in the presence of unit nonresponse

Let us start by considering a synthetic estimator using the [7,8] approach.

Lemma: Supposed that preference is given to the groups as a powerful factor in explaining the individual variation of elements within groups g's $(\mathrm{g} = 1,2,\ldots,\mathrm{G})$ considered being homogeneous for small area d's $(\mathrm{d} = 1,2,\ldots,\mathrm{D})$ under stratified sampling. Let the groups assumed as response homogeneity groups (RHG's), g's $(\mathrm{g} = 1,2,\ldots,\mathrm{G})$ be similar for small area d's $(\mathrm{d} = 1,2,\ldots,\mathrm{D})$, under stratified sampling. Then, (5) could be a modified calibration (synthetic) estimator for population mean in small area in the presence of unit nonresponse using the [7,8] procedure as follows:

i. $\hat{\bar{y}}_{drCU}^{*} = \sum_{r.g}^{G} W_{dg} \vartheta_{kU} \bar{y}_{.g}$ *(When information from the auxiliary variable is available at the population level), where* $\vartheta_{kU} = 1 + q_{dg}\left(\sum_{U.g}^{G} \bar{x}_{.g} - \sum_{r.g}^{G} W_{dg} \bar{x}_{.g}\right)\left(\sum_{r.g}^{G} W_{dg} q_{dg} \bar{x}_{.g}^2\right)^{-1} \bar{x}_{.g}$, for $g \in r_{.g}$

and,

ii. $\hat{\bar{y}}^{*}_{drCS} = \sum_{r.g}^{G} W_{dg}\vartheta_{ks}\bar{y}_{.g}$ *(When information from the auxiliary variable is only available at the sample level), where* $\vartheta_{ks} = 1 + q_{dg}\left(\sum_{s.g}^{G} W_{dg}\bar{x}_{.g} - \sum_{r.g}^{G} W_{dg}\bar{x}_{.g}\right)\left(\sum_{r.g}^{G} W_{dg}q_{dg}\bar{x}_{.g}^{2}\right)^{-1}\bar{x}_{.g}$

Proof: Let the calibration synthetic estimator $\hat{\bar{y}}^{*}_{dC}$ be given as

$$\hat{\bar{y}}^{*}_{dC} = \sum_{r.g}^{G} W^{*}_{dg}\bar{y}_{.g} \tag{5}$$

where $\bar{y}_{.g} = \sum_{r.d}^{D}\sum_{k\in r_{dg}}^{N_{dg}} \frac{y_{dgk}}{n_{.g}}$, $n_{.g} = \sum_{r.d}^{D} n_{dg}$ and W^{*}_{dg} the chosen calibration weights such that the chi-square type distance measure:

$$\Phi = \sum_{r.g}^{G} \frac{\left(W^{*}_{dg} - W_{dg}\right)^{2}}{W_{dg}q_{dg}} \tag{6}$$

is minimized, while satisfying the calibration constraints:

$$\sum_{r.g}^{G} W^{*}_{dg}\bar{x}_{.g} = \sum_{U.g}^{G} \bar{x}_{.g} \tag{7}$$

and

$$\sum_{r.g}^{G} W^{*}_{dg}\bar{x}_{.g} = \sum_{s.g}^{G} W_{dg}\bar{x}_{.g} \tag{8}$$

Case 1: Availability of information for auxiliary variable at the population level

Assume that information by the auxiliary variable is available at the population level, Info-U: then minimizing the distance function in (6) subject to the calibration constraint in (7) will give the optimization function;

$$\Phi(W, W^{*}) = \sum_{r.g}^{G} \frac{\left(W^{*}_{dg} - W_{dg}\right)^{2}}{W_{dg}q_{dg}} - 2\lambda\left[\sum_{r.g}^{G} W^{*}_{dg}\bar{x}_{.g} - \sum_{U.g}^{G} \bar{x}_{.g}\right]$$

After solving for the Lagrange multiplier λ, the calibration weight becomes;

$W^{*}_{dg} = W_{dg}\vartheta_{kU}$ (9)

substituting (9) in (5) will give

$\hat{\bar{y}}^{*}_{drCU} = \sum_{r.g}^{G} W_{dg}\vartheta_{kU}\bar{y}_{.g}$ (10)

where ϑ_{kU} is as earlier defined $\vartheta_{kU} = 1 + q_{dg}\left(\sum_{U.g}^{G} \bar{x}_{.g} - \sum_{r.g}^{G} W_{dg}\bar{x}_{.g}\right)\left(\sum_{r.g}^{G} W_{dg}q_{dg}\bar{x}_{.g}^{2}\right)^{-1}\bar{x}_{.g}$

Case 2: Non-availability of information for the auxiliary at the population level of the domain

Suppose that there is no information on the population mean of the auxiliary variable in the domain, calibration can be done on the unbiased estimate $\sum_{s.g} W_{dg}\bar{x}_{.g}$. Here, minimizing the distance measure in (6) subject to the calibration constraint in (8) will result in the optimization problem

$$\Phi(W, W^{*}) = \sum_{r.g}^{G} \frac{\left(W^{*}_{dg} - W_{dg}\right)^{2}}{W_{dg}q_{dg}} - 2\lambda\left[\sum_{r.g}^{G} W^{*}_{dg}\bar{x}_{.g} - \sum_{s.g}^{G} W_{dg}\bar{x}_{.g}\right]$$

which is solved, and the calibration weight obtained as

$$W^{*}_{dg} = W_{dg}\vartheta_{ks} \tag{11}$$

Substituting (11) in (5) will result in a new estimator under Info-S as follows:

$$\hat{\bar{y}}^{*}_{drCS} = \sum_{r.g}^{G} W_{dg}\vartheta_{ks}\bar{y}_{.g} \tag{12}$$

where ϑ_{sk} is as earlier defined as $\vartheta_{ks} = 1 + q_{dg}\left(\sum_{s.g}^{G} W_{dg}\bar{x}_{.g} - \sum_{r.g}^{G} W_{dg}\bar{x}_{.g}\right)\left(\sum_{r.g}^{G} W_{dg}q_{dg}\bar{x}_{.g}^{2}\right)^{-1}\bar{x}_{.g}$

Note: Although the estimators in equations (10) and (12) are useful in areas where there are small/no sample sizes, they are biased when an area of interest is characterized by small sample size and nonresponse.

2.3 New calibration estimator with alternative weights for small area in the presence of nonresponse

Here, we proposed a new estimator with an alternative distance measure and a new design weight d_k^* (which is the product of the original design weight $\frac{1}{\pi_k}$ and the inverse of the response probability φ_k) for the estimation of population mean $\bar{Y}_d$ to resolve the challenges of biasness and higher mean square error due to small sample size and nonresponse in small area estimation.

Proposition: Let the estimator of a population mean for small area in the presence of nonresponse be $\hat{\bar{y}}_{\mathrm{dr}} = \frac{1}{N_d}\sum_{g\in r.g}^{G}\sum_{k\in r_{dg}}^{N_{dg}}\frac{y_{dgk}}{\pi_k\hat{\theta}_k}$, then, an alternative estimator $\hat{\bar{y}}_{\mathrm{dr}}^{\mathrm{o}} = \bar{X}_d\hat{\mathrm{B}}_{\mathrm{drc}}$ that can produce a more reliable estimate under stratified random sampling can be obtained by defining a new design weight and calibrating on an alternative distance measure, $\sum_{r.g}^{G}\frac{\left(W_{dg}^{o}-W_{dg}\right)^{2}}{W_{dg}(W_{dg}-1)}$.

Proof: Recall the response probability θ_k in equation (1), and let the inverse of its estimate be given as $\varphi_k = \frac{1}{\hat{\theta}_k} = \frac{n_{dg}}{n_{r_{dg}}}$, then one can obtain a new design weight under nonresponse for the distance minimization as:

$$d_k^* = \frac{\varphi_k}{\pi_k} = \frac{N_{dg}}{n_{r_{dg}}} \tag{13}$$

and under stratified sampling, equation (1) can be written as:

$$\hat{\bar{y}}_{dr}^{*} = \frac{1}{N_d}\sum_{g\in r.g}^{G}\sum_{k\in r_{dg}}^{N_{dg}} d_k^* y_{dgk}$$

$$\hat{\bar{y}}_{dr}^{*} = \sum_{g\in r.g}^{G} W_{dg}\bar{y}_{.g} \tag{14}$$

Thus, the estimator for the population mean using calibration approach is given as:

$$\hat{\bar{y}}_{dr}^{0} = \sum_{g\in r.g}^{G} W_{dg}^{0}\bar{y}_{.g} \tag{15}$$

where W_{dg}^{o} is the chosen calibration weight such that the distance function as in [23] under stratified sampling

$$\Phi = \sum_{r.g}^{G}\frac{\left(W_{dg}^{o}-W_{dg}\right)^{2}}{W_{dg}(W_{dg}-1)} \tag{16}$$

is minimized subject to the calibration constraint:

$$\sum_{g\in r.g}^{G} W_{dg}^{0}\bar{y}_{.g} = \sum_{g\in s.g}^{G} W_{dg}\bar{x}_{.g} \tag{17}$$

and the optimization problem

$$\Phi(W, W^{0}) = \sum_{r.g}^{G}\frac{\left(W_{dg}^{o} - W_{dg}\right)^{2}}{W_{dg}\left(W_{dg} - 1\right)} - 2\lambda\left[\sum_{g\in r.g}^{G} W_{dg}^{0}\bar{y}_{.g} - \sum_{g\in s.g}^{G} W_{dg}\bar{x}_{.g}\right]$$

is solved for λ, and the calibration weights obtained as:

$$W_{dg}^{o} = W_{dg}\bar{x}_{.g}^{2}\left(\sum_{g\in r_{.g}}^{G} W_{dg}\bar{x}_{.g}^{2}\right)^{-1}\bar{X}_{d} \tag{18}$$

given that $\sum_{g\in s_{.g}}^{G} W_{dg}\bar{x}_{.g} = \bar{X}_{d}$.

Substituting (18) in (15) gives a new calibration estimator for small area in the presence of nonresponse as

$$\hat{\bar{y}}_{dr}^{o} = \bar{X}_{d}\hat{B}_{drc} \tag{19}$$

Equation (19) takes the form of the global regression-synthetic estimator of the population mean for small area, (see, for example, [6], where

$$\hat{B}_{drc} = \frac{\sum_{g\in r_{.g}}^{G} W_{dg}\bar{x}_{.g}\bar{y}_{.g}}{\sum_{g\in r_{.g}}^{G} W_{dg}\bar{x}_{.g}^{2}} \tag{20}$$

3 Results and Discussion

In this section, empirical investigation is carried out using simulation analysis in R. The procedure of population generation and sample selection for different sample settings in the simulation analysis is adopted from [22] and the response probability model is adopted from [23] as discussed in section 2.0. However, three different probability distributions are considered, namely; gamma, normal, and exponential distributions to depict different real-life scenarios.

3.1. Findings

The summary of the representation of units in each group across the domains and the results of the evaluation under nonresponse are obtained using simulated data as earlier discussed in section 2.0. Here, the theoretical formulations are validated and an indisputable pathway to the progress of small area estimation established. The results of the validation are as presented and discussed in Tables 1-3.

3.2. Discussion

Table 1 shows how the population was split into three groups with the respective values of intercepts and slopes for gamma, normal and exponential distributions. Table 2 presents the population of a broad domain under study divided into sub-domains and further partitioned into groups that are larger than the domains but cut across the domains to form grids that are mutually exclusive and exhaustive. Each simulation run in Table 2 involves the selection of $R = 100{,}000$, using R software for independent samples and the computation of various estimates for sample of sizes $n = 248(5\%)$, $n = 495(10\%)$, $n = 990(20\%)$, and $n = 1239(25\%)$ drawn using SRSWOR from U.

Table 1. Population groups with slopes and intercepts for different distributions

Distributions		**Gamma**		**Normal and Exponential**	
GROUP (g)	**Cells in groups**	$\boldsymbol{\beta_{0g}}$	$\boldsymbol{\beta_{1g}}$	$\boldsymbol{\beta_{0g}}$	$\boldsymbol{\beta_{1g}}$
1	U_{d1} for $k = 1,2,\ldots,10$	200	30	5	1.5
2	U_{d2} for $k = 11,12,\ldots,20$	300	20	10	2.5
3	U_{d3} for $k = 21,22,\ldots,30$	400	10	15	3.5

Table 2. Summary representation of units in each group and domains

Domains (d)	Groups (g)			Domains
	1	2	3	(U_d)
1	U_{11}	U_{12}	U_{13}	U_1
2	U_{21}	U_{22}	U_{23}	U_2
3	U_{31}	U_{32}	U_{33}	U_3
4	U_{41}	U_{42}	U_{43}	U_4
5	U_{51}	U_{52}	U_{53}	U_5
6	U_{61}	U_{62}	U_{63}	U_6
7	U_{71}	U_{72}	U_{73}	U_7
8	U_{81}	U_{82}	U_{83}	U_8
9	U_{91}	U_{92}	U_{93}	U_9
10	U_{101}	U_{102}	U_{103}	U_{10}
Groups $(U_{.g})$	$U_{.1}$	$U_{.2}$	$U_{.3}$	U

Table 3. $\overline{MSE}$, $\overline{\%ARB}$ and $\overline{\%CV}$ for Gamma (10,1), Norm (5,1) and Exp (0.5) with average response probabilities

SAM In %	DIS	$\hat{\bar{\theta}}$ in %	$\overline{\%ARB}$			$\overline{MSE}$			$\overline{\%CV}$		
			$\hat{\bar{y}}_{ds}$	$\hat{\bar{y}}^{*}_{drCS}$	$\hat{\bar{y}}^{0}_{dr}$	$\hat{\bar{y}}_{ds}$	$\hat{\bar{y}}^{*}_{drCS}$	$\hat{\bar{y}}^{0}_{dr}$	$\hat{\bar{y}}_{ds}$	$\hat{\bar{y}}^{*}_{drCS}$	$\hat{\bar{y}}^{0}_{dr}$
5	Gam	86	1148.8	58.7	0.2	3.3×10^{12}	8.6×10^{9}	1.1×10^{5}	1149.5	58.7	2.8
		70	1149.3	58.7	0.2	3.3×10^{12}	8.6×10^{9}	1.5×10^{5}	1149.9	58.7	2.9
		60	1148.7	58.7	0.2	3.3×10^{12}	8.6×10^{9}	1.0×10^{5}	1149.3	58.7	2.8
	Norm	86	1147.8	65.0	13.9	9.3×10^{9}	3.0×10^{7}	1.4×10^{6}	1148.3	65.0	14.0
		70	1149.5	65.1	13.9	9.3×10^{9}	3.0×10^{7}	1.4×10^{6}	1149.9	65.1	14.0
		60	1146.1	65.1	14.0	9.2×10^{9}	3.0×10^{7}	1.4×10^{6}	1146.6	65.1	14.2
	Exp	86	1145.5	65.7	16.5	4.2×10^{9}	1.4×10^{7}	8.7×10^{5}	1151.8	65.7	18.7
		70	1136.2	65.9	17.3	4.2×10^{9}	1.4×10^{7}	9.7×10^{5}	1142.5	65.9	19.5
		60	1137.1	65.8	16.9	4.2×10^{9}	1.4×10^{7}	9.2×10^{5}	1143.4	65.9	19.1
10	Gam	86	1148.6	58.7	0.2	3.3×10^{12}	8.6×10^{9}	1.3×10^{5}	1148.9	58.7	2.0
		70	1149.0	58.7	0.3	3.3×10^{12}	8.6×10^{9}	1.7×10^{5}	1149.3	58.7	2.0
		60	1148.7	58.7	0.2	3.3×10^{12}	8.6×10^{9}	1.5×10^{5}	1149.0	58.7	2.0
	Norm	86	1147.8	65.1	13.8	9.3×10^{9}	3.0×10^{7}	1.3×10^{6}	1148.0	65.1	13.9
		70	1149.5	65.1	13.9	9.3×10^{9}	3.0×10^{7}	1.4×10^{6}	1149.7	65.1	13.9
		60	1146.1	65.1	14.0	9.2×10^{9}	3.0×10^{7}	1.4×10^{6}	1146.3	65.1	14.0
	Exp	86	1142.5	65.7	16.4	4.2×10^{9}	1.4×10^{7}	8.7×10^{5}	1145.5	65.7	17.5
		70	1133.3	65.9	17.2	4.1×10^{9}	1.4×10^{7}	9.5×10^{5}	1136.3	66.0	18.3
		60	1133.6	65.9	16.8	4.1×10^{9}	1.4×10^{7}	9.1×10^{5}	1136.6	65.9	18.0
20	Gam	86	1148.5	58.7	0.2	3.3×10^{12}	8.6×10^{9}	1.5×10^{5}	1148.7	58.7	1.3
		70	1148.9	58.7	0.3	3.3×10^{12}	8.6×10^{9}	1.9×10^{5}	1149.0	58.7	1.4
		60	1148.6	58.7	0.3	3.3×10^{12}	8.6×10^{9}	1.8×10^{5}	1148.7	58.7	1.3
	Norm	86	1147.6	65.1	13.8	9.3×10^{9}	3.0×10^{7}	1.3×10^{6}	1147.7	65.1	13.9
		70	1149.4	65.1	13.9	9.3×10^{9}	3.0×10^{7}	1.4×10^{6}	1149.4	65.1	13.9
		60	1146.0	65.1	14.0	9.2×10^{9}	3.0×10^{7}	1.4×10^{6}	1146.1	65.1	14.0
	Exp	86	1141.1	65.7	16.3	4.2×10^{9}	1.4×10^{7}	8.6×10^{5}	1142.5	65.7	16.8
		70	1131.9	66.0	17.2	4.1×10^{9}	1.4×10^{7}	9.5×10^{5}	1133.2	66.0	17.7
		60	1132.3	65.9	16.8	4.1×10^{9}	1.4×10^{7}	9.1×10^{5}	1133.7	65.9	17.3
25	Gam	86	1148.5	58.7	0.3	3.3×10^{12}	8.6×10^{9}	1.7×10^{5}	1148.6	58.7	1.2
		70	1148.8	58.7	0.3	3.3×10^{12}	8.6×10^{9}	2.1×10^{5}	1148.9	58.7	1.2
		60	1148.4	58.7	0.3	3.3×10^{12}	8.6×10^{9}	1.7×10^{5}	1148.5	58.7	1.2
	Norm	86	1147.6	65.1	13.8	9.3×10^{9}	3.0×10^{7}	1.3×10^{6}	1147.7	65.1	13.9
		70	1149.4	65.1	13.9	9.3×10^{9}	3.0×10^{7}	1.4×10^{6}	1149.4	65.1	13.9
		60	1146.0	65.1	14.0	9.2×10^{9}	3.0×10^{7}	1.4×10^{6}	1146.1	65.1	14.0

SAM In %	DIS	$\bar{\hat{\theta}}$ in %	$\overline{\%ARB}$			$\overline{MSE}$			$\overline{\%CV}$		
			$\hat{\bar{y}}_{ds}$	$\hat{\bar{y}}^{*}_{drCS}$	$\hat{\bar{y}}^{0}_{dr}$	$\hat{\bar{y}}_{ds}$	$\hat{\bar{y}}^{*}_{drCS}$	$\hat{\bar{y}}^{0}_{dr}$	$\hat{\bar{y}}_{ds}$	$\hat{\bar{y}}^{*}_{drCS}$	$\hat{\bar{y}}^{0}_{dr}$
	Exp	86	1140.7	65.7	16.3	4.2×10^{9}	1.4×10^{7}	8.6×10^{5}	1141.7	65.7	16.7
		70	1131.6	65.9	17.2	4.1×10^{9}	1.4×10^{7}	9.5×10^{5}	1133.0	66.0	17.5
		60	1132.0	65.9	16.8	4.1×10^{9}	1.4×10^{7}	9.1×10^{5}	1133.0	65.9	17.2

Table 3 presents the results of the evaluation under nonresponse. As shown in column 6, the percent average absolute relative bias, $\overline{\%ARB}$ of the alternative calibration estimator $\hat{\bar{y}}^{0}_{dr}$, was negligible and almost unbiased with gamma distribution and considerably small in all sample settings for normal and exponential distributions, making it a more reliable calibration estimator for small area. These values are better than that produced by the calibration synthetic estimator $\hat{\bar{y}}^{*}_{drCS}$, and negligible compared to that of the existing estimator $\hat{\bar{y}}_{ds}$. The performance of the new calibration estimator is seen as an improvement of the [23] alternative distance measure in stratified sampling over that of [7] and [8]. As expected, this result agrees with the argument of [23] on the choice of weights in the presence of nonresponse. In addition, the effective reduction in the bias of $\hat{\bar{y}}^{0}_{dr}$ further justifies the suggestion by [21] on addressing both small area problems and nonresponse adjustment.

Again, in column 9 of Table 3, the result clearly shows that the new estimator, $\hat{\bar{y}}^{0}_{dr}$ was consistently more efficient under the three probability distributions and sample settings considered than $\hat{\bar{y}}^{*}_{drCS}$ and $\hat{\bar{y}}_{ds}$. Other than weight adjustments, this result is in tandem with the suggestions of [4] that partitioning the elements perceived to belong to the response homogeneous groups (RHGs) helps in reducing the variance of the interest variable.

Furthermore, the result of column 12 of Table 3 shows that $\hat{\bar{y}}^{0}_{dr}$ has $\overline{\%CV}$ between 1.2% to 2.9% for gamma distribution (which is less than 10% and makes the bias almost negligible), 13.9% to 19.1% for normal and exponential distributions for different sample settings, better than that of $\hat{\bar{y}}^{*}_{drCS}$ and $\hat{\bar{y}}_{ds}$. These values fall within the benchmark of 25% proposed by [24] for small area estimators and has given preference to $\hat{\bar{y}}^{0}_{dr}$ as being very suitable for small area estimation in the presence of nonresponse.

It is worthy of note that the result indicates a very strong correlation between units in the domains within groups. However, the result was different in-between the groups (strata), as the correlation was very weak with the elements of the population as informed by the simulation study. In fact, in the entire population, there was no significant correlation recorded for the three probability distributions, which supports the reason for stratification. Hence, the homogeneity created within the groups has paid off for both small sample size and nonresponse.

4 Conclusion

In conclusion, the concept of calibration on new design weights and alternative distance measure is a contribution towards the development of an ultimate estimator for small area estimation in the presence of unit nonresponse. This concept has yielded a fruitful result compared to the [7] and [8] approach. Consequently, this result will further enhance the disaggregation of national data and effective estimation in small areas with negligible biases in areas where there are small/no sample observations for proper policy implementations. More so, the New Calibration Estimator will be a useful tool in the hands of researchers and users of statistics by bridging the gap created because of small/no sample size and nonresponse in estimation of population parameters (mean or total) in small areas. Unlike the Statistical software, ETOS, (that takes care of adjusting the initial stratified simple random sample weight with the response set and then calibrate), Software developers should think towards developing a statistical software that will handle both the cross-sectional borrowing of strength using the stratified sampling design with the response set, and then calibrate with the alternative distance measure. As this will encourage wide application and usage.

5 Implication to Research

This paper considered the use of a calibration weighting scheme in small area estimation under stratified sampling design to produce reliable synthetic estimators of population mean in the presence of nonresponse. The paper also presented a calibration estimator with an alternative weighting scheme that exhibits smaller

relative biases, gain in efficiencies, and highly preferred coefficient of variations suitable for small area estimation. This supports the idea of calibration technique and weights adjustments using the assumed constraint on synthetic estimators under stratified sampling design for greatly improving the precision of estimators even in areas where there are smaller/no sample observations. The result of the coefficient of variation also suggests that when nonresponse occurs, it corresponds to an additional phase of sampling determined by the original sample design in line with [21] and the adoption of the alternative distance measure by [23] under stratified sampling has paid off. In terms of the probability distributions, the proposed estimator is more consistent in performance with gamma distribution preferably because of the choice of parameters, which may require further investigations in future work.

Emphatically, the use of this technique will drastically not only reduce sampling errors but will also minimize non-sampling errors, which in most cases seriously distort results of the survey as seen in most small area estimations. It therefore suffices to say that the proposed estimator has an endearing advantage over existing estimators of its class in SAE in the disaggregation of macro data with minimized error for planning and policy implementation in local areas.

Acknowledgement

We sincerely express our gratitude to the anonymous reviewers who painstakingly went through this manuscript with their constructive comments to make it a valid document for contribution to knowledge.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] National Center for Health Statistics. Synthetic state estimates of disability. Public Health Service Publication. Washington: U. S. Government Printing Office. 1968;1759.

[2] Gonzales ME. Use and evaluation of synthetic estimation. Proceedings of the American Statistical Association, Social Statistics Section. 1973;33-36.

[3] Sarndal CE. When robust estimation is not an obvious answer: The case of the synthetic estimator versus alternatives for small areas. Proceedings of the American Statistical Association, Survey Research Section. 1981;53-59.

[4] Sarndal CE, Swensson B, Wretman J. Model-assisted surveys. New York: Springer-Verlag; 1992.

[5] Marker DS. Organization of small area estimation using generalized linear regression framework. Journal of Official Statistics. 1999;15(1):1-24.

[6] Rao JNK. Small area estimation. New York: John Wiley; 2003.

[7] Lundstrom S, Sarndal CE. Calibration as a standard method for treating nonresponse. Journal of Official Statistics. 1999;15(2):305-327.

[8] Lundstrom S, Sarndal CE. Estimation in the presence of nonresponse and frame imperfections. Statistics Sweden; 2001.
ISBN: 91-618-1107-6.

[9] Sarndal CE, Lundstrom S. Estimation in survey with nonresponse. New York: John Wiley; 2005.

[10] Sarndal CE, Lundstrom S. Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. Journal of Official Statistics. 2008;24:167-191.

[11] Sarndal CE. The calibration approach in survey theory and practice. Survey Methodology. 2007;33:99-119.

[12] Lehtonen R, Sarndal CE, Veijanen A. The effect of model choice in estimation for domains, including small domains. Survey Methodology. 2003;1:33-44.

[13] Kott PS. Using calibration weighting to adjust for nonresponse and coverage errors. Survey Methodology. 2006;32:133-142.

[14] Lehtonen R, Veijanen A. Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics. 2012;66:125-133.

[15] Lehtonen R, Veijanen A. Small area estimation by calibration methods. World Statistics Conference, Rio de Janeiro; STS080; 2015.

[16] Pfeffermann D. New important developments in small area estimation. Statistical Sciences. 2013;28:40-68.

[17] Rota BJ, Laitila T. Comparison of some weighting methods for non-response adjustment. Lithuanian Journal of Statistics. 2015;54(1):69-83.

[18] Rao JNK, Molina I. Small area. 2nd ed. New York: John Wiley; 2015.
DOI: 10.1002/9781118735855.

[19] Rota BJ. Calibration adjustment for nonresponse in sample surveys. 2016.
Accessed 25 February 2018.
Available: www.oru.se/publikationer-avhandlingar

[20] Deville JC, Sarndal CE. Calibration estimators in survey sampling. Journal of the American Statistical Association. 1992;87:376-382.

[21] Guisti C, Rocco E. Small area estimation in the presence of nonresponse; 2013.
Accessed 29 December 2017.
Available:www.ds.unifi.it

[22] Hidiroglou MA, Estevao VM. A comparison of small area and calibration estimators via simulation. Statistics in Transition, New Series. 2014;17(1):133-154.

[23] Anderson PG. Optimal calibration weights under unit nonresponse in survey sampling. Survey Methodology. 2019;45(3):533-542.

[24] Molina I, Rao JNK. Small area estimation poverty indicators. Canadian Journal of Statistics. 2010;38:369-385.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/67176